An Empirical Study of Indirect Cross-validation

Olga Y. Savchuk, Jeffrey D. Hart, Simon J. Sheather

Abstract

In this paper we provide insight into the empirical properties of indirect cross-validation (ICV), a new method of bandwidth selection for kernel density estimators. First, we describe the method and report on the theoretical results used to develop a practical-purpose model for certain ICV parameters. Next, we provide a detailed description of a numerical study which shows that the ICV method usually outperforms least squares cross-validation (LSCV) in finite samples. One of the major advantages of ICV is its increased stability compared to LSCV. Two real data examples show the benefit of using both ICV and a local version of ICV.

KEY WORDS: Cross-validation; Bandwidth selection; Kernel density estimation, Integrated Squared Error, Mean Integrated Squared Error.

1 Introduction

Let $X_1, \ldots, X_n$ be a random sample from an unknown density f. A kernel density estimator of f at the point x is defined as

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\Big(\frac{x - X_i}{h}\Big), \tag{1}$$

where $h > 0$ is the bandwidth, and K is the kernel, which is generally chosen to be a unimodal probability density function that is symmetric about zero and has finite variance. A popular choice for K is the Gaussian kernel: $\phi(u) = (2\pi)^{-1/2} \exp(-u^2/2)$. To distinguish between estimators with different kernels, we shall refer to estimator (1) with given kernel K as a *K-kernel estimator*.

Practical implementation of the estimator (1) requires specification of the smoothing parameter h. The two most widely used bandwidth selection methods are least squares cross-validation, proposed independently by Rudemo (1982) and Bowman (1984), and the Sheather and Jones (19 plug-in method. Plug-in is often preferred since it produces more stable bandwidths than does LSCV. Nevertheless, the LSCV method is still popular since it requires fewer assumptions than the plug-in method and works well when the density is difficult to estimate; see Loader (1999), van Es (1992), and Sain, Baggerly, and Scott (1994).

The main flaw of LSCV is high variability of the selected bandwidths. Other drawbacks include the tendency of cross-validation curves to exhibit multiple local minima with the first local minimum being too small (see Hall and Marron (1991)), and the tendency of LSCV to select bandwidths that are much too small when the data exhibit a small amount of autocorrelation (see Hart and Vieu (1990) and Cao and Vilar Fernandez (1993) for results of a numerical study). Many modifications of LSCV have been proposed in an attempt to improve its performance. These include biased cross-validation of Scott and Terrell (1987), a method of Chiu (1991), the trimmed cross-validation of Feluch and Koronacki (1992), the modified cross-validation of Stute (1992), and the method of Ahmad and Ran (2004) based on kernel contrasts.

This paper is concerned with a new modification of the LSCV method, called *indirect cross-validation* (ICV), recently proposed by the authors Savchuk, Hart, and Sheather (2008). The ICV method depends on two parameters, α and σ. A main theoretical result is that at asymptotically optimal choices of α and σ the ICV bandwidth can converge to zero at a rate $n^{-1/4}$, which is substantially better than the $n^{-1/10}$ rate of LSCV. The present paper contains the results of an empirical study of ICV. In Section 2 we provide a description of the method. Section 3 contains the details underlying the development of a practical purpose model for α and σ. Section 4 outlines the results of a numerical study which, in particular, show that ICV has greater stability in finite samples than does LSCV. In Section 5 we apply ICV and a local version of ICV to real data sets. Section 6 provides a summary of our results.

2 Description of indirect cross-validation

2.1 Notation and definitions

We begin with some notation and definitions that will be used subsequently. For an arbitrary function g, define

$$R(g) = \int g(u)^2\,du, \quad \mu_{jg} = \int u^j g(u)\,du,$$

where here and subsequently integrals are assumed to be over the whole real line. The popular measures of performance of the kernel estimators (1) are integrated squared error (ISE) and mean integrated squared error (MISE). The ISE is defined as

$$ISE(h) = \int \left(\hat{f}_h(x) - f(x)\right)^2 dx, \tag{2}$$

and MISE is defined as the expectation of ISE. Assuming that the underlying density f has second derivative which is continuous and square integrable and that $R(K) < \infty$, the bandwidth which asymptotically minimizes the MISE of the $K-$kernel estimator (1) has the following form:

$$h_n = \left\{ \frac{R(K)}{\mu_{2K}^2 R(f'')} \right\}^{1/5} n^{-1/5}. \tag{3}$$

The LSCV criterion is given by

$$LSCV(h) = R(\hat{f}_h) - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{h,-i}(X_i), \tag{4}$$

where $\hat{f}_{h,-i}$ denotes the kernel estimator (1) constructed from the data without the observation X_i. A well known fact is that $LSCV(h)$ is an unbiased estimator of $MISE(h) - \int f^2(x)\,dx$. For this reason the LSCV method is often called *unbiased cross-validation*. Let $\hat{h}_{UCV}$ and h_0 denote the bandwidths which minimize the LSCV function (4) and the MISE of the ϕ-kernel estimator. Section 2.2 defines the ICV bandwidth, denoted as $\hat{h}_{ICV}$.

2.2 The basic method

The essence of the ICV method is to use different kernels at the cross-validation and density estimation stages. The same idea is exploited by the one-sided cross-validation method of Hart and Yi (1998) in the regression context. ICV first selects the bandwidth of an $L-$kernel estimator using least squares cross-validation. Selection kernels L used for this purpose are described in Section 2.3. The bandwidth so obtained is rescaled so that it can

be used with the ϕ-kernel estimator. The multiplicative constant C has the following form:

$$C = \left(\frac{\mu_{2L}^2}{2\sqrt{\pi} R(L)^2} \right)^{1/5}, \tag{5}$$

which is motivated by the asymptotically optimal MISE bandwidth (3).

2.3 Selection kernels

We consider the family of kernels $\mathcal{L} = \{L(\,\cdot\,;\alpha,\sigma) : \alpha \geq 0, \sigma > 0\}$, where, for all u,

$$L(u;\alpha,\sigma) = (1+\alpha)\phi(u) - \frac{\alpha}{\sigma}\phi\left(\frac{u}{\sigma}\right). \tag{6}$$

Note that the Gaussian kernel is a special case of (6) when $\alpha = 0$ or $\sigma = 1$. Each member of $\mathcal{L}$ is symmetric about 0 and has the second moment $\mu_{2L} = \int u^2 L(u)\,du = 1+\alpha-\alpha\sigma^2$. It follows that kernels in $\mathcal{L}$ are second order, with the exception of those for which $\sigma = \sqrt{(1+\alpha)/\alpha}$.

The family $\mathcal{L}$ can be partitioned into three families: $\mathcal{L}_1$, $\mathcal{L}_2$ and $\mathcal{L}_3$. The first of these is $\mathcal{L}_1 = \left\{L(\cdot;\alpha,\sigma) : \alpha > 0, \sigma < \frac{\alpha}{1+\alpha}\right\}$. Each kernel in $\mathcal{L}_1$ has a negative dip centered at $x = 0$. The kernels in $\mathcal{L}_1$ are ones that "cut-out-the-middle," some examples of which are shown in Figure 1**(a)**.

The second family is $\mathcal{L}_2 = \left\{L(\cdot;\alpha,\sigma) : \alpha > 0, \frac{\alpha}{1+\alpha} \leq \sigma \leq 1\right\}$. Kernels in $\mathcal{L}_2$ are densities which can be unimodal or bimodal. Note that the Gaussian kernel is a member of this family. The third family is $\mathcal{L}_3 = \left\{L(\cdot;\alpha,\sigma) : \alpha > 0, \sigma > 1\right\}$, each member of which has negative tails. Examples are shown in Figure 1**(b)**.

Kernels in $\mathcal{L}_1$ and $\mathcal{L}_3$ turn out to be highly efficient for cross-validation purposes but very inefficient for estimating f. This explains why we do not use L as both a selection *and* an estimation kernel.

Selection kernels in $\mathcal{L}$ are mixtures of two normal densities, which greatly simplifies computations. In particular, closed form expressions exist for the $LSCV$ and ISE functions. This fact has been utilized by Marron and Wand (1992) to derive exact MISE expressions. Marron and Wand (1992) point out that, in addition to their computational advantages, normal mixtures can approximate any density arbitrarily well in various senses. Mixtures of normals are therefore an excellent model for use in simulation studies, a fact which we take advantage of in Section 4.

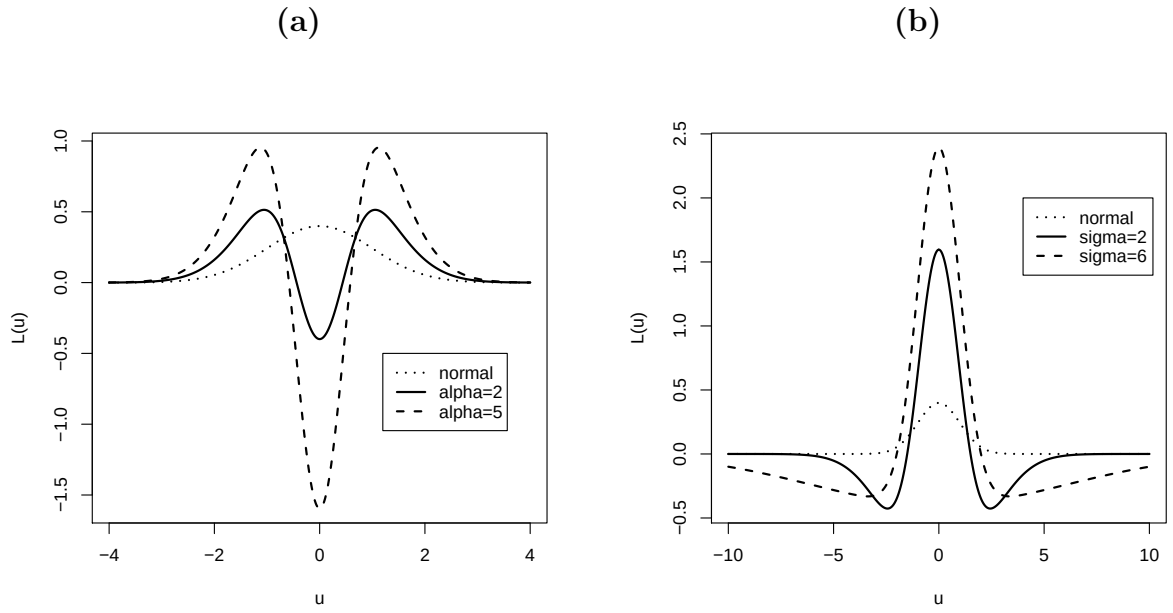

Figure 1: **(a)** Selection kernels in $\mathcal{L}_1$ which have $\sigma = 0.5$; **(b)** Selection kernels in $\mathcal{L}_3$ with $\alpha = 6$. The dotted curve in both graphs corresponds to the Gaussian kernel.

3 Practical issues

In this section we address the problem of choosing the parameters, α and σ, of the selection kernel in practice. We review some large sample theory for the ICV method and provide the theoretical results used to develop the practical-purpose model for α and σ.

3.1 Large sample theory

Large sample theory was developed in Savchuk, Hart, and Sheather (2008) by considering the asymptotic mean squared error (MSE) of the ICV bandwidth. Their results may be summarized as follows.

1. Under suitable regularity conditions the ICV bandwidth is asymptotically normally distributed.

2. The asymptotic MSE of $\hat{h}_{ICV}$ has been found for two cases: $\sigma \to 0$ (cut-out-the-middle kernels) and $\sigma \to \infty$ (negative-tailed kernels). It turns out that when the asymptotically optimal values of α and σ are used in the respective cases, the MSE converges to zero at the same rate of $n^{-9/10}$, but the limiting ratio of optimum mean

squared errors is 0.752, with $\sigma \to \infty$ yielding the smaller error. In comparison, the rate at which the MSE for $\hat{h}_{UCV}$ converges to zero is $n^{-6/10}$.

The subsequent theoretical results are provided for the case $\sigma \to \infty$.

3. The relative rate of convergence of $\hat{h}_{ICV}$ to h_0 is $n^{-1/4}$, whereas the corresponding rate for $\hat{h}_{UCV}$ is $n^{-1/10}$.

4. Values of σ which minimize the asymptotic MSE are as follows:

$$\sigma_{n,opt} = n^{3/8} A_\alpha \left[\frac{R(f)R(f'')^{13/5}}{R(f''')^2} \right]^{5/8}, \tag{7}$$

where $A_\alpha = 16\sqrt{\pi}\frac{2^{7/16}}{3^{5/8}}\frac{\alpha^{3/4}}{(1+\alpha)^2}\left(\frac{1}{8}(1+\alpha)^2 - \frac{8}{9\sqrt{3}}(1+\alpha) + \frac{1}{\sqrt{2}}\right)^{5/8}$.

5. The asymptotically optimal α is 2.4233. Remarkably, the optimal α does not depend on f.

6. When the asymptotically optimal values of α and σ are used, the asymptotic bias and standard deviation of $\hat{h}_{ICV}$ converge to zero at the same rate of $n^{-9/20}$.

3.2 MSE-optimal α and σ

Asymptotic results are not always reliable for practical purposes. In order to have an idea of whether the negative-tailed or cut-out-the middle kernels should really be used, and how good choices of α and σ vary with n and f, we considered the following expression for the asymptotic MSE of the ICV bandwidth:

$$\begin{aligned} \mathrm{MSE}(\hat{h}_{ICV}) = \; & \left(\frac{1}{4\pi}\right)^{1/5} \frac{R(f''')^2}{R(f'')^{16/5}} n^{-3/5} \left\{ \frac{2}{25} \frac{R(f)R(f'')^{13/5}}{R(f''')^2} \frac{R(\rho_L)}{R(L)^{9/5}(\mu_{2L}^2)^{1/5}} + \right. \\ & \left. \frac{n^{-3/5}}{400} \left(\frac{R(L)^{2/5}\mu_{2L}\mu_{4L}}{(\mu_{2L}^2)^{7/5}} - \frac{3}{(4\pi)^{1/5}} \right)^2 \right\}. \end{aligned} \tag{8}$$

Expression (8) is valid for either large or small values of σ and includes second order bias terms.

As our target densities we considered the following five normal mixtures defined in the article by Marron and Wand (1992):

	Density									
	normal		skewed unimodal		bimodal		separated bimodal		skewed bimodal	
n	α	σ	α	σ	α	σ	α	σ	α	σ
100	3.05	2.79	5.28	1.68	109.68	1.03	16.70	1.19	343.74	1.01
250	2.78	4.04	3.16	2.60	48.46	1.06	4.51	1.84	177.15	1.02
500	2.73	4.97	2.84	3.56	6.21	1.55	3.18	2.58	161.39	1.02
1000	2.69	5.97	2.75	4.49	3.73	2.12	2.84	3.54	123.78	1.03
5000	2.61	8.84	2.66	6.85	2.77	4.26	2.70	5.74	4.71	1.79
20000	2.55	12.40	2.59	9.58	2.68	6.22	2.63	8.08	2.85	3.46
100000	2.50	18.80	2.53	14.27	2.60	9.19	2.56	11.94	2.70	5.65
500000	2.47	29.54	2.49	21.88	2.54	13.65	2.50	18.07	2.62	8.39

Table 1: MSE-optimal α and σ.

Gaussian density:	$N(0,1)$
Skewed unimodal density:	$\frac{1}{5}N(0,1) + \frac{1}{5}N\left(\frac{1}{2}, \left(\frac{2}{3}\right)^2\right) + \frac{3}{5}N\left(\frac{13}{12}, \left(\frac{5}{9}\right)^2\right)$
Bimodal density:	$\frac{1}{2}N\left(-1, \left(\frac{2}{3}\right)^2\right) + \frac{1}{2}N\left(1, \left(\frac{2}{3}\right)^2\right)$
Separated bimodal density:	$\frac{1}{2}N\left(-\frac{3}{2}, \left(\frac{1}{2}\right)^2\right) + \frac{1}{2}N\left(\frac{3}{2}, \left(\frac{1}{2}\right)^2\right)$
Skewed bimodal density:	$\frac{3}{4}N(0,1) + \frac{1}{4}N\left(\frac{3}{2}, \left(\frac{1}{3}\right)^2\right)$.

These choices for f represent density shapes that are common in practice.

In Table 1 we provide the MSE-optimal choices of α and σ for the target densities at eight sample sizes ranging from $n = 100$ up to $n = 500000$. It is obvious that the MSE-optimal α and σ vary greatly from one density to another, which is especially true for "small" sample sizes. However, the optimal α seems to converge to about 2.5 for each density as n increases, which fits with our observation that the optimal α is 2.4233. The optimal σ is increasing with sample size. It us remarkable that all the MSE-optimal α and σ in Table 1 correspond to kernels from $\mathcal{L}_3$, the family of negative-tailed kernels.

3.3 Model for the ICV parameters

We found a practical purpose model for α and σ by using polynomial regression. Our independent variable was $\log_{10}(n)$ and our dependent variables were the MSE-optimal values of $\log_{10}(\alpha)$ and $\log_{10}(\sigma)$ for different densities. The $\log_{10}$ transformations for α and σ stabilize

n	100	250	500	1000	5000	20000	100000	500000
α_{mod}	25.20	12.77	8.24	5.71	3.23	2.66	2.66	2.62
σ_{mod}	1.39	1.89	2.37	2.95	4.83	7.21	11.22	16.98

Table 2: Model choices of α and σ.

variability. Using a sixth degree polynomial for α and a quadratic for σ, we arrived at the following models for α and σ:

$$\begin{aligned} \alpha_{mod} &= 10^{3.390-1.093\log 10(n)+0.025\log 10(n)^3-0.00004\log 10(n)^6} \\ \sigma_{mod} &= 10^{-0.58+0.386\log 10(n)-0.012\log 10(n)^2}, \end{aligned} \tag{9}$$

which are appropriate for $100 \leq n \leq 500000$. The MSE-optimal values of $\log_{10}(\alpha)$ and σ together with the model fits are shown in Figure 2. In Table 2 we give the model choices α_{mod} and σ_{mod} for the same sample sizes as in Table 1.

4 Simulation study

The primary goal of our simulation study was to compare ICV with ordinary LSCV. However, we will also provide simulation results for the Sheather-Jones plug-in method.

We considered the four sample sizes $n = 100$, 250, 500 and 5000, and took samples from the target densities listed in Section 3.2. For each combination of density and sample size we did 1000 replications. In all cases the parameters α and σ in the selection kernel L were chosen according to model (9).

Let $\hat{h}_0$ denote the minimizer of $ISE(h)$ for a Gaussian kernel estimator. For each sample, we computed $\hat{h}_0$, $\hat{h}^*_{ICV}$, $\hat{h}_{UCV}$ and the Sheather-Jones plug-in bandwidth $\hat{h}_{SJPI}$. The definition of $\hat{h}^*_{ICV}$ is as follows:

$$\hat{h}^*_{ICV} = \min(\hat{h}_{ICV}, \hat{h}_{OS}), \tag{10}$$

where $\hat{h}_{OS}$ is the oversmoothed bandwidth of Terrell (1990). It is arguable that *no* data-driven bandwidth should be larger than $\hat{h}_{OS}$ since this statistic estimates an upper bound for *all* MISE-optimal bandwidths (under standard smoothness conditions).

For any random variable Y defined in each replication of our simulation, we denote the mean, standard deviation and median of Y over all replications (with n and f fixed) by $\widehat{\mathrm{E}}(Y)$, $\widehat{\mathrm{SD}}(Y)$ and $\widehat{\mathrm{Median}}(Y)$. To evaluate the bandwidth selectors we computed $\widehat{\mathrm{E}}\big\{ISE(\hat{h})/ISE(\hat{h}_0)\big\}$

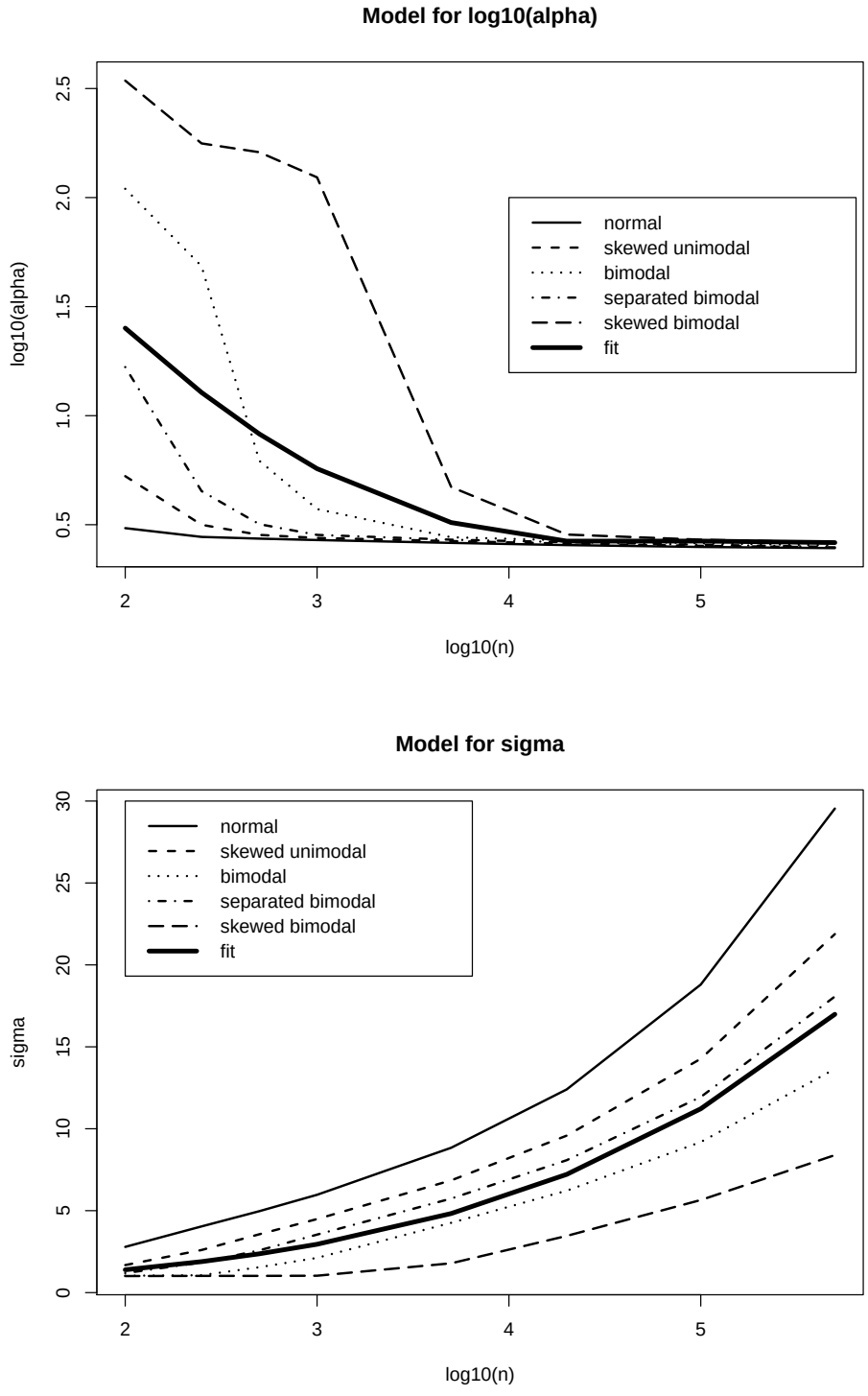

Figure 2: MSE-optimal $\log_{10}(\alpha)$ and σ and the model fits.

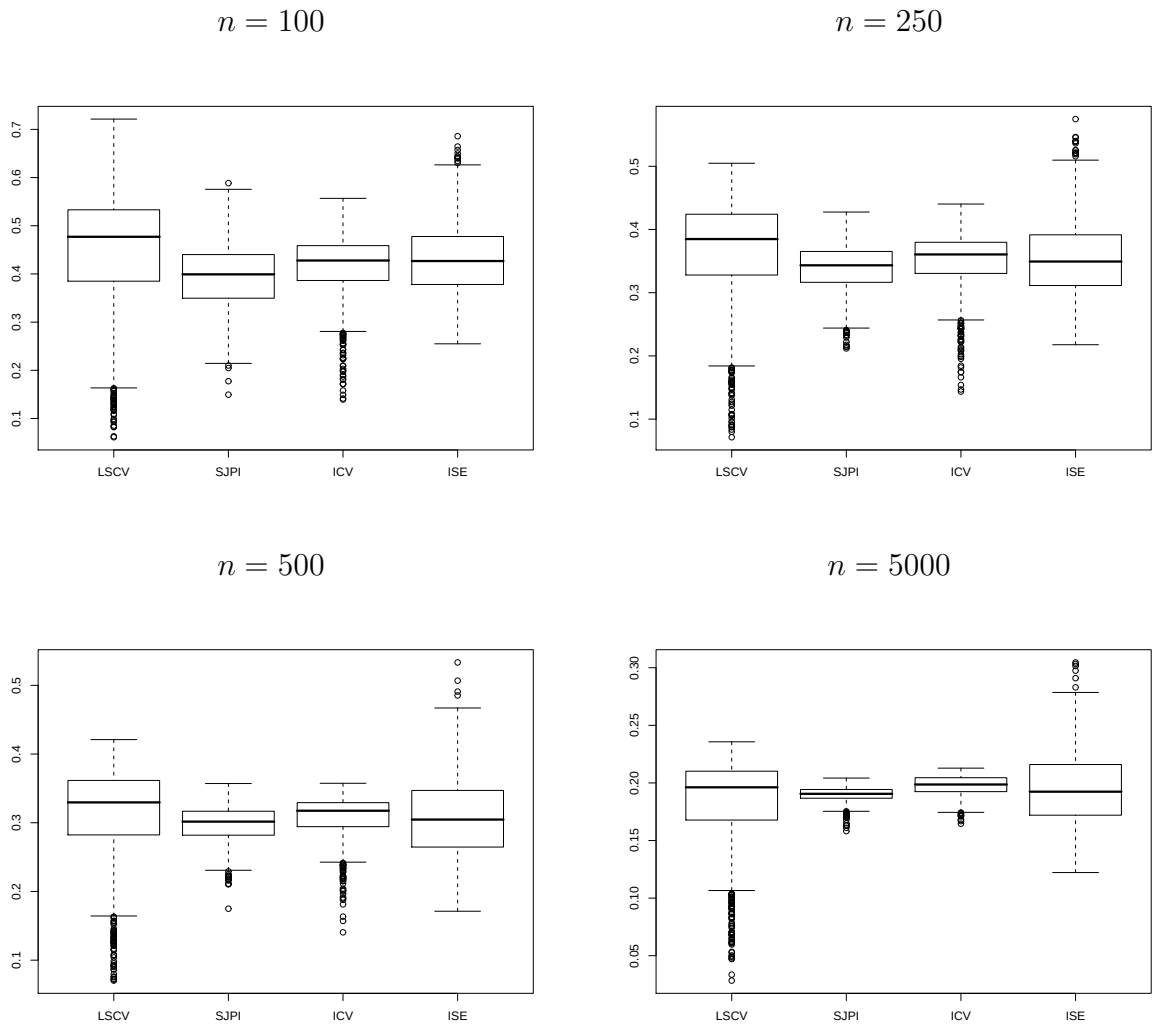

Figure 3: Boxplots for the data-driven bandwidths in case of the **Normal density**.

and $\widehat{\text{Median}}\{ISE(\hat{h})/ISE(\hat{h}_0)\}$ for $\hat{h}$ equal to each of $\hat{h}^*_{ICV}$, $\hat{h}_{UCV}$ and $\hat{h}_{SJPI}$. We also computed the performance measure $\widehat{E}\left(\hat{h} - \hat{E}(\hat{h}_0)\right)^2$, which estimates the MSE of the bandwidth $\hat{h}$.

Our main simulation results for the "normal" and "bimodal" densities, as defined in Section 3.2, are given in Tables 3 and 4 and Figures 3 and 4. Results for the other densities are available from the authors. Other statistics reported in Tables 3 and 4 are $\widehat{\text{E}}(\hat{h})$ and $\widehat{\text{SD}}(\hat{h})$ for each type of bandwidth considered.

The reduced variability of the ICV bandwidth is evident in our study. The ratio $\widehat{\text{SD}}(\hat{h}^*_{ICV})/\widehat{\text{SD}}(\hat{h}_{UCV})$ ranged between 0.9713 and 0.2103 in the twenty settings considered. However, the variances of the ICV bandwidths were always higher compared to the Sheather-Jones plug-in band-

n	**LSCV**	**SJPI**	**ICV**	**ISE**
$\widehat{\mathrm{E}}(\hat{h})$				
100	0.44524596	0.39338747	0.41530230	0.43162318
250	0.36398008	0.33883538	0.34944737	0.35487029
500	0.31094126	0.29803205	0.30864570	0.30806146
5000	0.18359629	0.18992356	0.19768683	0.19526358
$\widehat{\mathrm{SD}}(\hat{h}) \cdot 10^2$				
100	12.32173263	6.43244579	6.52298637	7.52008697
250	8.35772162	3.71742374	4.44775700	6.27300326
500	7.11168918	2.60300987	3.08015801	5.63495059
5000	3.90077096	0.61900268	0.82041632	3.09277421
$\widehat{\mathrm{E}}(\hat{h} - \widehat{\mathrm{E}}(\hat{h}_0))^2 \cdot 10^4$				
100	153.52907115	55.95467615	45.17051435	
250	70.61154173	16.37660421	20.05684094	
500	50.60847941	7.77477660	9.48129936	
5000	16.56205491	0.66793916	0.73113122	
$\widehat{\mathrm{E}}\big(\mathrm{ISE}(\hat{h})/\mathrm{ISE}(\hat{h}_0)\big)$				
100	2.46997542	1.90795915	1.72178966	
250	1.91593730	1.50563016	1.47567596	
500	1.75806058	1.37734003	1.36096679	
5000	1.41316047	1.11460567	1.10313807	
$\widehat{\mathrm{Median}}\big(\mathrm{ISE}(\hat{h})/\mathrm{ISE}(\hat{h}_0)\big)$				
100	1.31108630	1.15695876	1.11233574	
250	1.21715835	1.10408948	1.09365380	
500	1.21396609	1.10306404	1.09608944	
5000	1.10907960	1.04471055	1.05183075	

Table 3: Simulation results for the **Gaussian density**.

n	**LSCV**	**SJPI**	**ICV**	**ISE**
$\widehat{\mathrm{E}}(\hat{h})$				
100	0.42908686	0.39453431	0.41955286	0.38237337
250	0.31360942	0.31160054	0.32846189	0.29715278
500	0.25927533	0.26238646	0.27450416	0.25320682
5000	0.15262210	0.15706804	0.16255246	0.15478049
$\widehat{\mathrm{SD}}(\hat{h}) \cdot 10^2$				
100	13.56532316	7.44425312	9.56680379	7.60899932
250	8.46734473	4.18778288	6.50918853	4.29431763
500	5.70587208	2.44443305	4.20078840	3.55982408
5000	2.46293965	0.47951752	0.81457083	1.96503777
$\widehat{\mathrm{E}}(\hat{h} - \widehat{\mathrm{E}}(\hat{h}_0))^2 \cdot 10^4$				
100	205.65547766	56.84037076	105.25535253	
250	74.33244070	19.60736507	52.12977298	
500	32.89268754	6.81193546	22.16474597	
5000	6.10659189	0.28203637	1.26689717	
$\widehat{\mathrm{E}}\big(\mathrm{ISE}(\hat{h})/\mathrm{ISE}(\hat{h}_0)\big)$				
100	1.69951929	1.32733595	1.36143018	
250	1.51599857	1.20914143	1.28743335	
500	1.41670996	1.15070890	1.19168891	
5000	2.06430484	1.06839987	1.07675906	
$\widehat{\mathrm{Median}}\big(\mathrm{ISE}(\hat{h})/\mathrm{ISE}(\hat{h}_0)\big)$				
100	1.20951575	1.08744161	1.13356965	
250	1.16087896	1.08338970	1.12699702	
500	1.12243694	1.06072702	1.09421867	
5000	1.05825025	1.03067963	1.03649944	

Table 4: Simulation results for the **Bimodal density**.

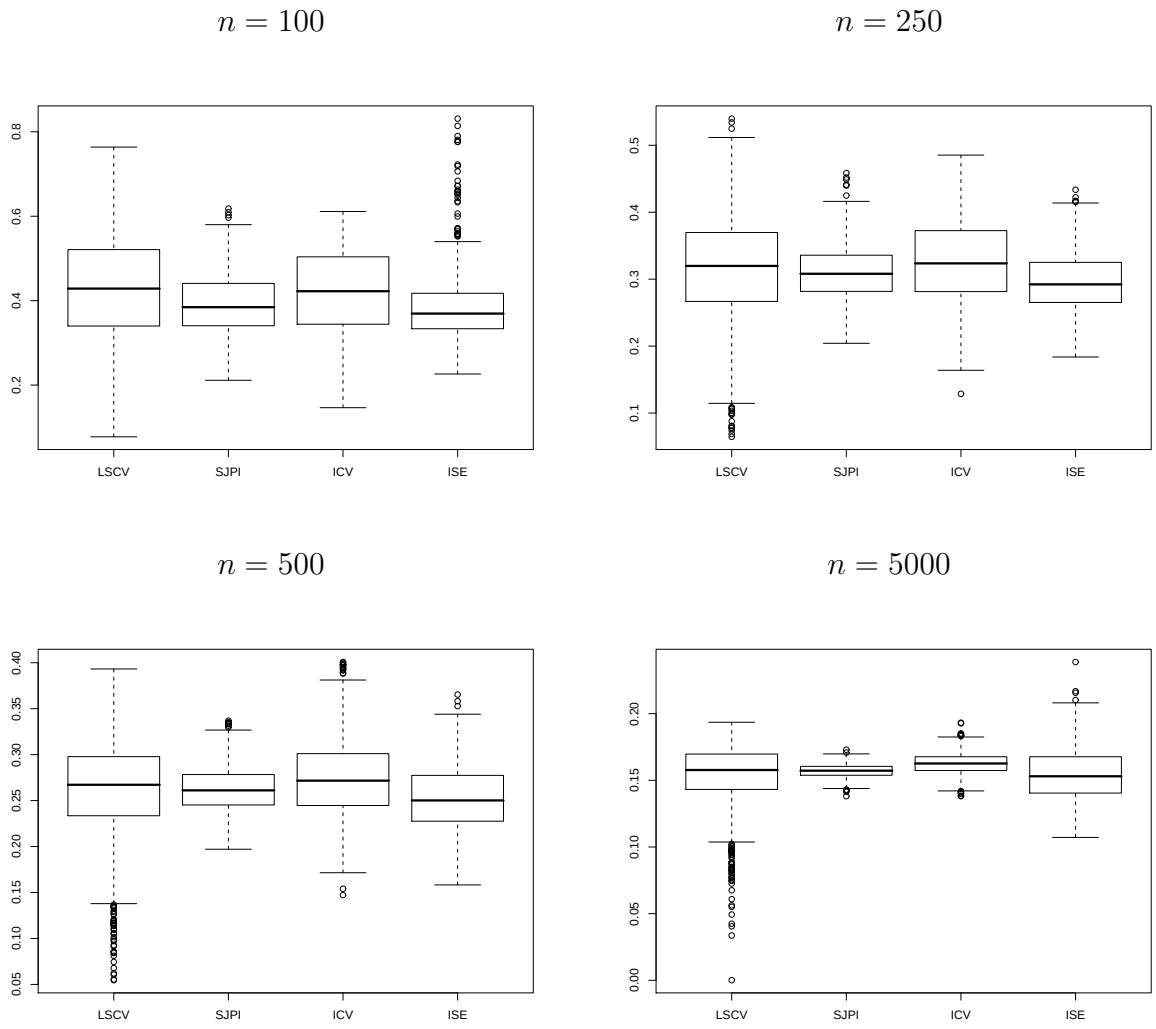

Figure 4: Boxplots for the data-driven bandwidths in case of the **Bimodal density**.

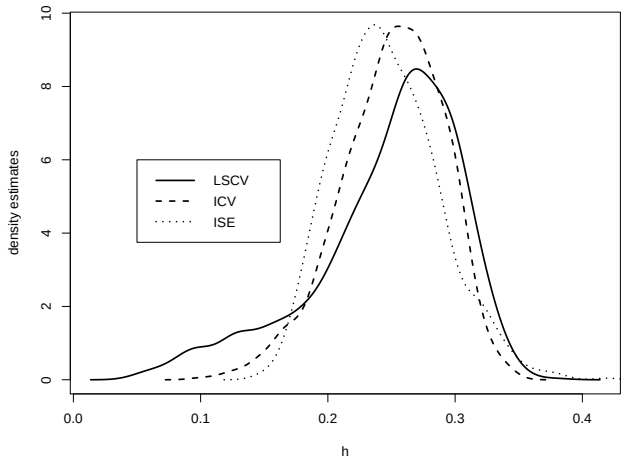

Figure 5: Kernel density estimates for random bandwidths from the simulation with the Skewed Unimodal density and $n = 250$.

widths. It is worth noting that the ratio of sample standard deviations of the ICV and LSCV bandwidths decreases as the sample size n increases.

The mean squared distance $\widehat{\mathrm{E}}\left(\hat{h} - \widehat{\mathrm{E}}(\hat{h}_0)\right)^2$ was smaller for the ICV method than for the LSCV method in all but two cases corresponding to the Skewed Bimodal density, $n = 250$ and 500. Plug-in always had a smaller value of $\widehat{\mathrm{E}}\left(\hat{h} - \widehat{\mathrm{E}}(\hat{h}_0)\right)^2$ than did ICV.

The most important observation is that the values of $\widehat{\mathrm{E}}\big(ISE(\hat{h})/ISE(\hat{h}_0)\big)$ were smaller for ICV than for LSCV for all combinations of densities and sample sizes. The values of $\widehat{\mathrm{Median}}\big(ISE(\hat{h})/ISE(\hat{h}_0)\big)$ were smaller for ICV than for LSCV in all but one case, which corresponds to the Skewed Bimodal density at $n = 250$ when $\widehat{\mathrm{Median}}\big(ISE(\hat{h}_{ICV})/ISE(\hat{h}_0)\big)$ was 1.0013 times greater than $\widehat{\mathrm{Median}}\big(ISE(\hat{h}_{UCV})/ISE(\hat{h}_0)\big)$.

Despite the fact that the LSCV bandwidth is asymptotically normally distributed (see Hall and Marron its distribution in finite samples tends to be skewed to the left. In our simulations we have noticed that the distribution of the ICV bandwidth is less skewed than that of the LSCV bandwidth. A typical case is illustrated in Figure 5, where kernel density estimates for the two data-driven bandwidths are plotted from the simulation with the Skewed Unimodal density at $n = 250$. Also plotted is a density estimate for the ISE-optimal bandwidths. Note that the ICV density is more concentrated near the middle of the ISE-optimal distribution than the density estimate for LSCV.

Figure 6 provides scatterplots of the bandwidths $\hat{h}_{UCV}$ and $\hat{h}_{ICV}$ versus $\hat{h}_0$ in the case

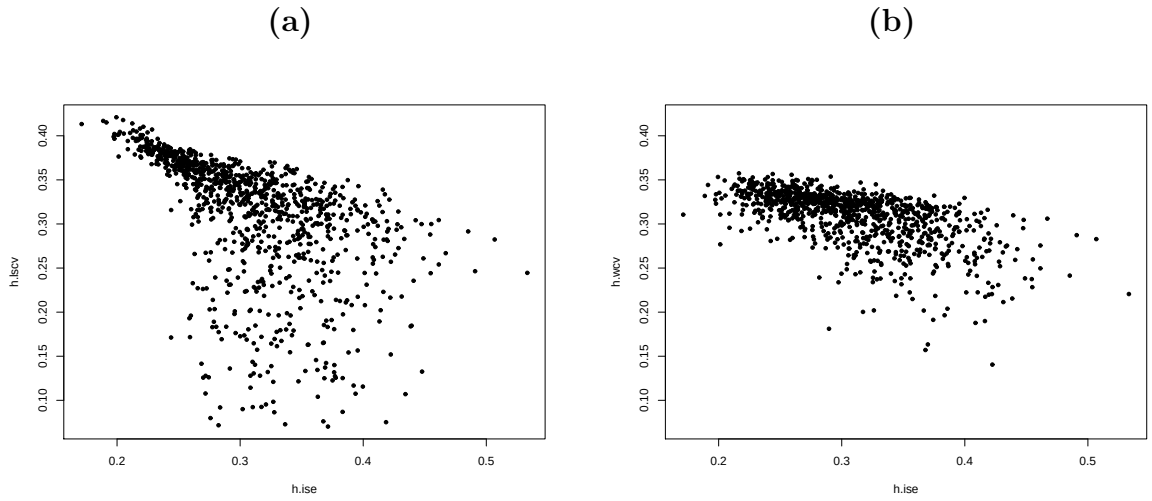

Figure 6: Scatterplots of $\hat{h}$ vs. $\hat{h}_0$ for the case of a Gaussian density and $n = 500$, with $\hat{h}$ corresponding to the **(a)** LSCV and **(b)** ICV bandwidths.

of the Gaussian density and $n = 500$. The sample correlation coefficients were -0.52 and -0.60 for LSCV and ICV, respectively. The fact that these correlations are negative is a well-established phenomenon; see, for example, Hall and Johnstone (1992). Note that the ICV bandwidths cluster more tightly about the MISE minimizer $h_0 = 0.315$.

A problem we have noticed with the ICV method is that its criterion function can have two local minima when the sample size is moderate and the density has two modes. The following example illustrates the problem. In Figure 7(a) we have plotted three ICV curves for the case of the Separated Bimodal density and $n = 100$. The minimizers of the solid, dashed and dotted lines occur at the h-values 0.2991, 2.0467 and 0.2204, respectively. For comparison, the corresponding bandwidths chosen by the Sheather-Jones plug-in method are 0.3240, 0.2508 and 0.2467. The value of $h = 2.0467$ which minimizes the dashed ICV curve is obviously too large. The local minimum at 0.1295 would yield a much more reasonable estimate. The problem of choosing too large a bandwidth from the second local minimum is mitigated by using the rule (10). Indeed, the oversmoothed bandwidths for the three samples are shown by the vertical lines in Figure 7 and were 0.7404, 0.7580 and 0.7341. Note that the problem with the ICV curve having two local minima of approximately the same value quickly goes away as the sample size increases. This is illustrated in Figure 7(b), where we have plotted three criterion curves for the Separated Bimodal case with $n = 500$. Thus, the selection rule $\hat{h}^*_{ICV}$ given by (10) rather than just $\hat{h}_{ICV}$ appears to be useful mostly for small and moderate sample sizes.

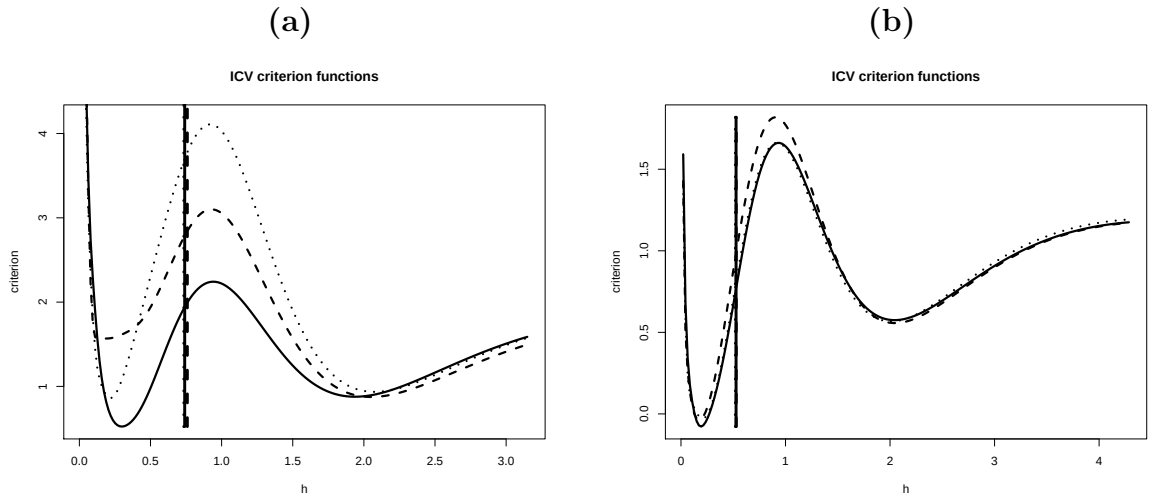

Figure 7: Three ICV criterion functions in case of the Separated Bimodal density at **(a)** $n = 100$ and **(b)** $n = 500$.

5 Real data examples

In this section we show how the ICV method works on two real data sets. The purpose of the first example is to compare the performance of the ICV, LSCV, and Sheather-Jones plug-in methods. The second example illustrates the benefit of using ICV locally.

5.1 PGA data

In this example the data are the average numbers of putts per round played, for the top 175 players on the 1980 and 2001 PGA golf tours. The question of interest is whether there has been any improvement from 1980 to 2001. This data set has already been analyzed by Sheather (2004) in the context of comparing the performances of LSCV and Sheather-Jones plug-in.

In Figure 8 we have plotted an unsmoothed frequency histogram and the LSCV, ICV and Sheather-Jones plug-in density estimates for a combined data set of 1980 and 2001 putting averages. The class interval size in the unsmoothed histogram was chosen to be 0.01, which corresponds to the accuracy to which the data have been reported. There is a clear indication of two modes in the histogram.

The estimate based on the LSCV bandwidth is apparently undersmoothed. The ICV and plug-in estimates look similar and have two modes, which agrees with evidence from the unsmoothed histogram and seems reasonable since the data were taken from two populations.

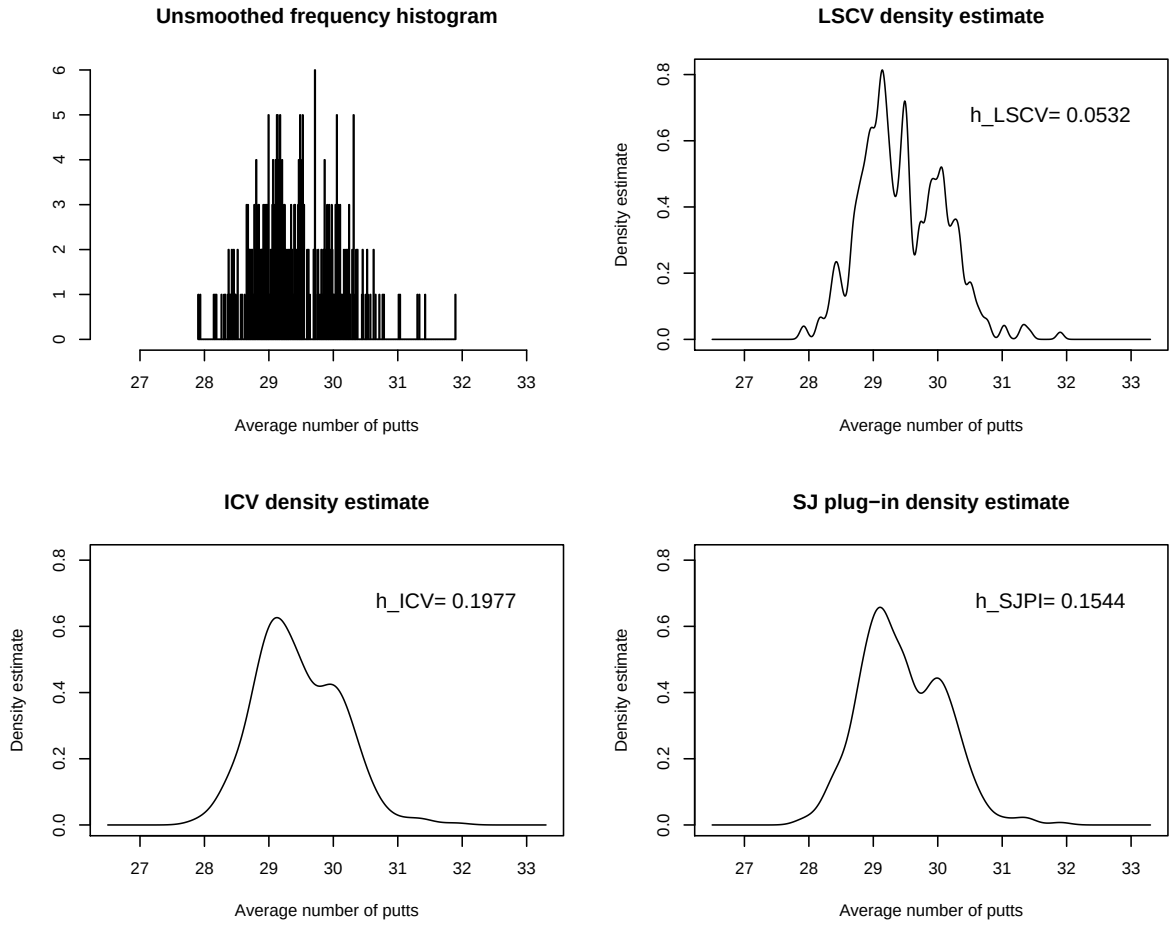

Figure 8: Unsmoothed frequency histogram and kernel density estimates for average numbers of putts per round from 1980 and 2001 combined.

In Figure 9 we have plotted kernel density estimates separately for the years 1980 and 2001. ICV seems to produce a reasonable estimate in both years, whereas LSCV yields a very wiggly and apparently undersmoothed estimate in 2001.

5.2 Local ICV example

Local cross-validation methods for density estimation, independently proposed by Hall and Schucany (1989) and Mielniczuk, Sarda, and Vieu (1989), consist in performing LSCV at each value of the argument x using a fraction of the data that are close to x. Allowing the bandwidth to depend on x is desirable when the smoothness of the underlying density changes sufficiently with x.

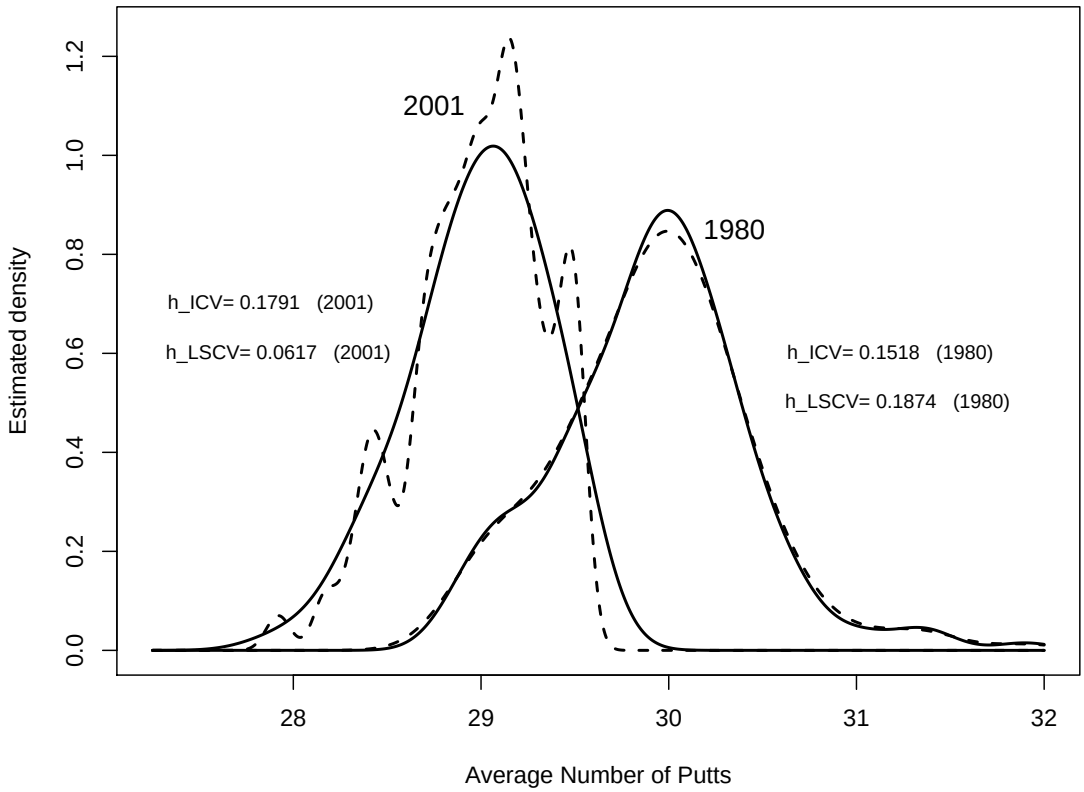

Figure 9: Kernel density estimates based on LSCV (dashed curve) and ICV (solid curve) produced separately for the data from 1980 and 2001.

The local ICV method was introduced in Savchuk, Hart, and Sheather (2008). It is different from the local LSCV method in that it uses ICV rather than LSCV for the local bandwidth selection. Another difference is that local ICV uses the first local minimizer of the local criterion function as opposed to the global minimizer of local LSCV.

The local ICV criterion function is defined as

$$ICV(x,b,w) = \frac{1}{w}\int \phi\left(\frac{x-u}{w}\right)\hat{f}_b(u)^2\,du - \frac{2}{nw}\sum_{i=1}^{n}\phi\left(\frac{x-X_i}{w}\right)\hat{f}_{b,-i}(X_i),$$

where function $\hat{f}_b$ is the kernel density estimate based on a selection kernel L with a smoothing parameter b. The quantity w defines the extent to which the cross-validation is local, with a large choice of w corresponding to global ICV. Let $\hat{b}(x)$ be the first local minimum of the local ICV curve for the fixed value of x. Then the corresponding bandwidth

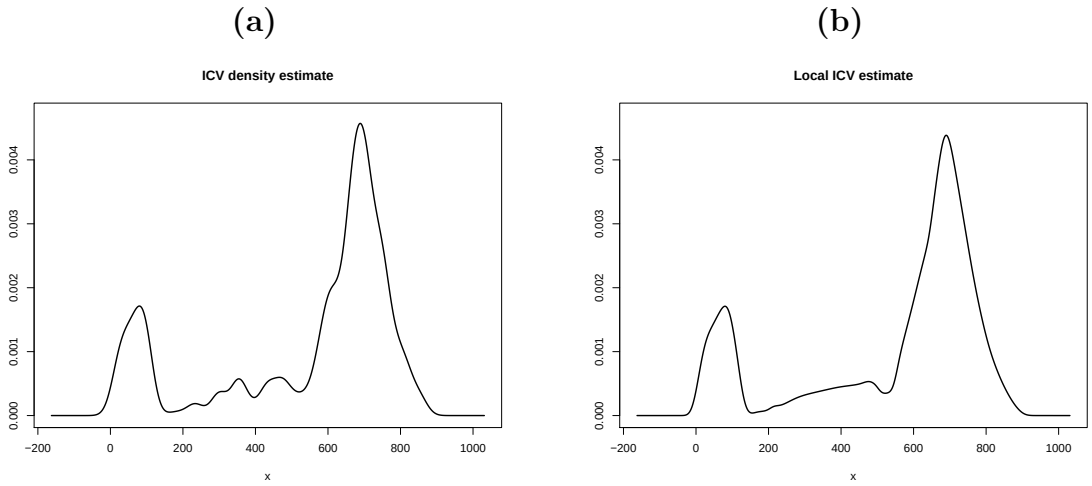

Figure 10: Density estimates for the DC data set with **(a)** being the global ICV density estimate and **(b)** corresponding to the local ICV estimate.

of a $\phi-$kernel estimator is defined as $\hat{h}(x) = C\hat{b}(x)$, where C is computed as in (5). Local ICV outperformed the local LSCV method in a simulated data example in the article of Savchuk, Hart, and Sheather (2008). In this paper we show how local ICV and LSCV perform in a real data example.

We analyze the data of size $n = 517$ on the Drought Code (DC) of the Canadian Forest Fire Weather index (FWI) system. DC is one of the explanatory variables which can be used to predict the burned area of a forest in the Forest Fires data set. This data can be downloaded from the website `http://archive.ics.uci.edu/ml/datasets/Forest+Fires`. The data were collected and analyzed by Cortez and Morais (2007).

We computed the LSCV, ICV and Sheather-Jones plug-in bandwidths for the DC data. The LSCV method failed by yielding $\hat{h}_{UCV} = 0$. ICV and Sheather-Jones plug-in bandwidths were very close and produced similar density estimates. Figure 10 **(a)** gives the ICV density estimate. It shows two major modes connected with a wiggly curve, which indicates that varying the bandwidth with x may yield a smoother estimate of the underlying density.

Local ICV and LSCV have been applied to the DC data. We used $w = 40$ for both methods and the selection kernel with $\alpha = 6$ and $\sigma = 6$ for local ICV. This (α, σ) choice performed quite well for unimodal densities in our simulation studies on global ICV, and hence seems to be reasonable for local bandwidth selection since locally the density should have relatively few features. Let $x_{(i)}$, $i = 1, \ldots, n$, denote the ith member of the ordered sequence of observations. The local ICV and LSCV bandwidth were found for 50 evenly

spaced points in the interval $x_{(1)} - 0.2(x_{(n)} - x_{(1)}) \leq x \leq x_{(n)} + 0.2(x_{(n)} - x_{(1)})$. It turns out that in 45 out of 50 cases the local LSCV curve tends to $-\infty$ as $h \to 0$, which implies that the local LSCV estimate can not be computed. All 50 local ICV bandwidths were positive. We found a smooth function $\hat{h}(x)$ by interpolating at other values of x via a spline. The corresponding local ICV estimate, given in Figure 10**(b)**, shows a smoother density estimate.

6 Summary

Indirect cross-validation is a method of bandwidth selection in the univariate kernel density estimation context. The method first selects the bandwidth of an $L-$kernel estimator by least squares cross-validation, and then rescales this bandwidth so that it is appropriate for use in a Gaussian kernel density estimator. Selection kernels L have the form $(1+\alpha)\phi(u)-\alpha\phi(u/\sigma)/\sigma$, where $\alpha \geq 0$, $\sigma > 0$ and ϕ is the Gaussian kernel. Optimal kernels from this class yield bandwidths with relative error that converges to 0 at a rate of $n^{-1/4}$, which is a substantial improvement over the $n^{-1/10}$ rate of LSCV.

A practical purpose model for the selection kernel parameters, α and σ, has been developed. The model was built by performing polynomial regression on the MSE-optimal values of $\log_{10}(\alpha)$ and $\log_{10}(\sigma)$ at different sample sizes for five target densities. Use of this model makes the ICV method completely automatic.

An extensive simulation study showed that in finite samples ICV is more stable than LSCV. Although both ICV and LSCV bandwidths are asymptotically normal, the distribution of the ICV bandwidth for finite n is usually more symmetric and better concentrated in the middle of the density for ISE-optimal bandwidths. Using an oversmoothed bandwidth as an upper bound for the bandwidth search interval reduces the bias of the method and prevents selecting an impractically large value of h when the criterion curves exhibit multiple local minima.

The ICV method performs well in real data examples. ICV applied locally yields density estimates which are more smooth than estimates based on a single bandwidth. Often, local ICV estimates may be found when the local LSCV estimates do not exist.

References

Ahmad, I. A. and I. S. Ran (2004). Kernel contrasts: a data-based method of choosing smoothing parameters in nonparametric density estimation. *J. Nonparametr.*

Stat. 16(5), 671–707.

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika 71*(2), 353–360.

Cao, R., Q. d. R. A. and J. Vilar Fernandez (1993). Bandwidth selection in nonparametric density estimation under dependence : a simulation study. *Computational Statistics 8*, 313– 332.

Chiu, S.-T. (1991). Bandwidth selection for kernel density estimation. *Ann. Statist. 19*(4), 1883–1905.

Cortez, P. and A. Morais (2007). A data mining approach to predict forest fires using meteorological data. *in J. Neves, M. F. Santos and J. Machado Eds., New Trends in Artificial Intelligence, Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence, December, Guimaraes, Portugal*, 512–523.

Feluch, W. and J. Koronacki (1992). A note on modified cross-validation in density estimation. *Comput. Statist. Data Anal. 13*(2), 143–151.

Hall, P. and I. Johnstone (1992). Empirical functional and efficient smoothing parameter selection. *J. Roy. Statist. Soc. Ser. B 54*(2), 475–530. With discussion and a reply by the authors.

Hall, P. and J. S. Marron (1987). Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation. *Probab. Theory Related Fields 74*(4), 567–581.

Hall, P. and J. S. Marron (1991). Local minima in cross-validation functions. *J. Roy. Statist. Soc. Ser. B 53*(1), 245–252.

Hall, P. and W. R. Schucany (1989). A local cross-validation algorithm. *Statist. Probab. Lett. 8*(2), 109–117.

Hart, J. D. and P. Vieu (1990). Data-driven bandwidth choice for density estimation based on dependent data. *Ann. Statist. 18*(2), 873–890.

Hart, J. D. and S. Yi (1998). One-sided cross-validation. *J. Amer. Statist. Assoc. 93*(442), 620–631.

Loader, C. R. (1999). Bandwidth selection: classical or plug-in? *Ann. Statist. 27*(2), 415–438.

Marron, J. S. and M. P. Wand (1992). Exact mean integrated squared error. *Ann. Statist. 20*(2), 712–736.

Mielniczuk, J., P. Sarda, and P. Vieu (1989). Local data-driven bandwidth choice for density estimation. *J. Statist. Plann. Inference 23*(1), 53–69.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scand. J. Statist. 9*(2), 65–78.

Sain, S. R., K. A. Baggerly, and D. W. Scott (1994). Cross-validation of multivariate densities. *J. Amer. Statist. Assoc. 89*(427), 807–817.

Savchuk, O. Y., J. D. Hart, and S. J. Sheather (2008). Indirect cross-validation for density estimation. *J. Amer. Statist. Assoc., submitted*.

Scott, D. W. and G. R. Terrell (1987). Biased and unbiased cross-validation in density estimation. *J. Amer. Statist. Assoc. 82*(400), 1131–1146.

Sheather, S. J. (2004). Density estimation. *Statist. Sci. 19*(4), 588–597.

Sheather, S. J. and M. C. Jones (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc. Ser. B 53*(3), 683–690.

Stute, W. (1992). Modified cross-validation in density estimation. *J. Statist. Plann. Inference 30*(3), 293–305.

Terrell, G. R. (1990). The maximal smoothing principle in density estimation. *J. Amer. Statist. Assoc. 85*(410), 470–477.

van Es, B. (1992). Asymptotics for least squares cross-validation bandwidths in nonsmooth cases. *Ann. Statist. 20*(3), 1647–1657.